САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

КАФЕДРА МОДЕЛИРОВАНИЯ ЭЛЕКТРОМЕХАНИЧЕСКИХ И КОМПЬЮТЕРНЫХ СИСТЕМ

Бебяков Александр Михайлович

Выпускная квалификационная работа магистра

Разработка алгоритма биоинформатического анализа идентификации патогенных микроорганизмов

Направление 01.04.02

Прикладная математика и информатика

Научный руководитель,
кандидат биол. наук,
доцент
Семёнов А. В.

Куратор,
кандидат физ.-мат. наук,
доцент
Никифоров К. А.

Санкт-Петербург

2018

Содержание

Введение

Некультивируемое состояние обнаружено у многих патогенных видов микроорганизмов. В связи с тем, что рутинные бактериологические методы неприменимы для обнаружения некультивируемых форм, развитию таких методик как анализ на основе секвенирования по методу дробовика следует уделить значительное внимание. Подобный метод позволит в кратчайшие сроки определять наличие возбудителей особо опасных инфекций и, таким образом, значительно ускорять процесс принятия решений о мерах противодействия возможным эпидемиям.

Методы NGS – методы секвенирования нового поколения являются высокопроизводительными методами определения нуклеотидных последовательностей. Наиболее распространённой технологией NGS является Illumina [1], позволяющая извлекать из подготовленного образца за один запуск продолжительностью до 30 часов около 400 млн генетических последовательностей длиной до 600 символов.

В связи с всеобъемлемостью присущей подобного рода исследованию требуется разработка систем, осуществляющих применение методов классификации, способных обеспечить проведение точного, полного и эффективного по времени определения таксономической принадлежности получаемых последовательностей.

Из известных и важных применений указанного анализа можно заметить отделение личных данных о генетике исследуемого человека, представляющих значительный объём данных результатов секвенирования образцов, что влечет также ускорение последующих этапов анализа. Разделение по таксономической принадлежности позволяет повысить скорость, специфичность и точность работы таких инструментов анализа биологических данных как программы сборки (объединения в более крупные последовательности по пересечениям) и аннотации (разметки участков нуклеотидных последовательностей по функциональным проявлениям) геномных данных.

Для решения данной задачи классификации применяются методы, основанные на прямых сравнениях (BLAST [2], BWA [3]), но, учитывая количество нуклеотидных последовательностей получаемых из образцов, содержащих смеси организмов, подобного рода анализ может занять непозволительно большое количество времени.

Другим подходом является анализ последовательностей на основе количественных характеристик строк. Методы, применяющие данный подход, способны быстро производить оценку принадлежности к биологическим группам. [4]

Целью данной работы является разработка алгоритма анализа, позволяющего в кратчайшие сроки определять таксономическую принадлежность строк представляющих биологические последовательности.

Предполагается применение методов, основанных на анализе количественных характеристик строк с применением технологий машинного обучения с учителем. [5]

Машинное обучение позволяет ускорять разработку стратегии анализа, производя определение скрытых правил на основе ряда статистических экспериментов. [6]

Постановка задачи

Рассматривается задача классификации биологических последовательностей, являющихся фрагментами геномов вирусов или бактерий и полученных в результате секвенирования, по таксономической принадлежности. Каждая из таких последовательностей представима в форме строки s ненулевой длины l_s над алфавитом $\Sigma = \{A, T, G, C\}$.

Пусть X — множество описаний объектов (всевозможных строк из символов {A, T, G, C}), Y — множество идентификаторов биологических групп. Существует неизвестная целевая зависимость — отображение $y^*: X \to Y$, значения которой известны только на объектах конечной обучающей выборки $X_m = \{(s_1, t_1), \dots, (s_m, t_m)\}$. Требуется построить алгоритм $g: X \to Y$, способный классифицировать произвольный объект s из множества X.

В ходе подобной классификации, строке s приписывается определенный таксономический идентификатор (tax_id) t_s таким образом, чтобы обеспечить наибольшее сходство с эталонной последовательностью s^* с идентификатором $t_{s^*} = t_s$ из обучающей выборки.

Под сходством в данной работе понимается взвешенное качество выравнивания [7], которое представляет собой вариант редакционного расстояния и вычисляется как максимальный элемент следующей матрицы H:

$$\begin{gathered} H(i, 0) = 0,\ 0 \le i \le m \\ H(0, j) = 0,\ 0 \le j \le n \\ H(i, j) = \max\{0,\ H(i-1, j-1) + z(s_i, s^*_j),\ \max_{k \ge 1}\{H(i-k, j) + w(k)\}, \\ \max_{l \ge 1}\{H(i, j-l) + w(l)\}\},\ 1 \le i \le q,\ 1 \le j \le n, \end{gathered} \tag{1}$$

где q - длина s, а n - длина s^*;

$z(a, b)$ – некоторая функция сходства символов a и b алфавита Σ, принимающая неотрицательные значения при $a = b$ и неположительные при $a \neq b$ (например, $z(a, b) = 1$ при $a = b$ и $z(a, b) = -2$ при $a \neq b$);

$w(k)$ – функция штрафа за вставку или удаление, принимает неположительные значения и задается некоторой невозрастающей

зависимостью от размера k необходимой вставки или удаления (например, $w(k) = -2k$).

В связи с особенностями строения биологических последовательностей (наличие повторяющихся сегментов, функциональные элементы характерные для множества представителей обширных таксономических групп) не всегда доступно точное определение видовой принадлежности последовательности, которая указывается для эталонных последовательностей таких баз, как NCBI [8]. Так как различные биологические виды могут иметь значительные различия, в том числе и в отношении патогенности для человека, не следует приписывать двусмысленным последовательностям случайную видовую принадлежность. Следовательно, требуется обеспечить возможность классифицировать принадлежность последовательностей и к более обширным группам.

При предположении, что существует последовательность s^{**} обучающей выборки с идентификатором $t_{s^{**}} \neq t_{s^{*}}$, и взвешенное качество выравнивания s на s^{**} отличается от взвешенного качества выравнивания s на s^{*} на значение меньшее заданного порога, идентификация последовательности s не считается допустимой.

Исходя из информации о включении таксономических групп, представленной таксономическим деревом (рис. 1) [9], возможно составление обучающих выборок для реализации алгоритмов определения принадлежности строки таксономической группе любого уровня вложенности.

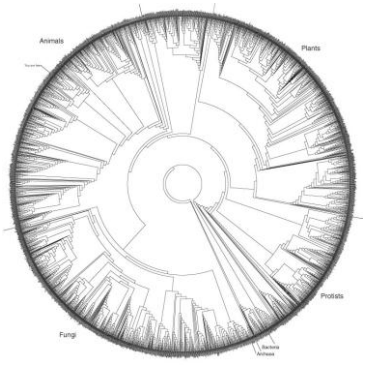

Рис. 1. Таксономическое дерево

Обзор литературы

В статье [5] представлены предположения о возможности применения множества различных методов классификации биологических последовательностей и проведен обзор множества статистических методов, методов выравнивания последовательностей (методы прямых сравнений) на последовательности из базы, такие как алгоритмы Нидлмана–Вунша, Смита–Ватермана, методы скрытых Марковских моделей. Авторами рассмотрены методы оценки качества результатов работы алгоритмов классификации и отмечена необходимость для решения биологических задач и задачи классификации в частности изучения таких техник, как теория информации, параллельные вычисления и машинное обучение.

Обзор алгоритмов идентификации бактерий сообществ на основе анализа рибосомальной ДНК представлен в статье [10], среди которых отмечена применимость методов скрытых Марковских моделей и наивного байесовского классификатора.

Большая устойчивость к генетическим изменениям и сокращение необходимых вычислений в задаче таксономической классификации организмов для методов основанных на анализе количественных характеристик строк отмечены в статье [4], в которой проведен обзор особенностей и методов обработки последовательностей вирусов.

Метод случайных лесов находит множество применений для различных задач классификации, таких как диагностика неисправности оборудования [11], подсчет участников дорожного движения [12], анализ изображений и анализ устойчивости ВИЧ к лекарственным препаратам [13].

В статье [14] происходит описание задач метагеномики и обзор методов извлечения данных образцов и их влияние на развитие области знаний о некультивируемых представителей микроорганизмов, в том числе микробиома человека (полный набор микроорганизмов присутствующих в человеческом теле, микрофлора). Авторы отмечают важность и доступность технологии секвенирования Illumina и дают представление о больших объемах данных

производимых метагеномными исследованиями, предполагающих необходимость разработки высокопроизводительных методов обработки и эффективных методов хранения результатов исследований.

Глава 1. Глубинный анализ данных

Для нахождения допустимой таксономической идентификации наибольшей глубины необходимо проведение классификации каждого объекта (строки) в следующей последовательности:

корень → домен (вирусы или бактерии),

вирусы → класс по Балтимору → семейство → род → вид → подвид,

бактерии → класс → семейство → род → вид → подвид.

Классификация объекта s начинается с присвоения ему таксономического идентификатора $t_s = 1$, соответствующего корню таксономического дерева. Если в ходе классификации s присваивается новый идентификатор нижележащего уровня, то алгоритм продолжает работу, пока не будет достигнут листовой узел дерева или последующий шаг не будет объявлен недопустимой идентификацией.

Для составления классификаторов [15] на основе машинного обучения с учителем для всех узлов каждого указанного ранга таксономического дерева необходимо выполнить этапы:

1. Сбор данных
2. Подготовка данных
3. Выбор модели анализа данных
4. Подбор параметров модели
5. Обучение модели
6. Анализ качества обучения

Детали выполнения перечисленных этапов обсуждаются в последующих главах.

Глава 2. Сбор данных

2.1. Необходимые данные

Для реализации алгоритмов на основе машинного обучения с учителем необходимо составить обучающее множество элементов вида объект-класс $X_m = \{(s_1, t_1), \dots, (s_m, t_m)\}$. В текущей задаче обучающее множество должно быть составлено из строковых представлений биологических последовательностей s_j, соответствующих бактериям и вирусам I-IV групп патогенности, представленных в Классификации биологических агентов, вызывающих болезни человека, по группам патогенности [16] с указанием таксономического идентификатора t_j.

Важную роль в систематизации живого мира имеет таксономическое дерево, являющееся формой представления связей «предок-потомок» присущей биологическим группам. Например, виды, принадлежащие одному роду, обозначаются как потомки вершины с идентификатором соответствующим указанному роду. Подобное знание позволит конструировать обучающие выборки для промежуточных вершин данного дерева и обеспечивать необходимую подготовку для построения иерархического классификатора. Также для лучшего представления о структуре таксономического дерева требуется указать для каждого таксономического идентификатора его ранг (уровень).

2.2. NCBI

Для сбора достоверных данных об искомых последовательностях необходимо обращение к известным биологическим базам данных.

Наиболее полная и подтвержденная информация о базах данных ДНК и РНК (GenBank, RefSeq), базах данных статей научной литературы и таксономической информации (Taxonomy) предоставляется Национальным центром биотехнологической информации (NCBI). [17]

RefSeq – база данных, находящаяся в открытом доступе и содержащая подтвержденные исследованиями эталонные последовательности геномов

более 70000 биологических видов. Данные последовательности предоставляются в формате fasta [18], содержащем последовательности символов алфавита Σ = {A, T, G, C}, разделенные заголовками последовательностей с информацией об идентификаторе в базе и именовании последовательности.

Taxonomy – база данных о классификации биологических последовательностей, содержащихся в том числе в базах данных GenBank и RefSeq.

RefSeq и Taxonomy доступны через систему Entrez, имеющую клиент-программу EDirect для командной строки UNIX. [19]

2.3. Автоматизация сбора данных.

Поиск таксономических идентификаторов соответствующих искомым биологическим видам (из списка СанПин) производился в базе Taxonomy с помощью EDirect. Загрузка необходимых последовательностей в формате fasta по обнаруженным идентификаторам и дочерних к ним, представляющих собой последовательности подвидов, если таковые были доступны, осуществлялась EDirect из базы данных RefSeq.

Построен индекс (файл taxid_index.tsv) необходимых tax_id. Биологические последовательности сгруппированы в файлы, именованные по таксономическому идентификатору в виде <tax_id>.fasta для каждого <tax_id> из taxid_index.tsv. Таким образом, была составлена база из 1130 последовательностей принадлежащих 438 различным биологическим группам (tax_id) патогенных вирусов и 585 последовательностей принадлежащих 164 различным tax_id патогенных бактерий.

Была загружена информация о таксономическом дереве из NCBI Taxonomy в виде двух файлов – nodes.dmp и names.dmp [8], содержащих соответственно информацию о связях «предок-потомок» для tax_id и информацию о ранге и научном именовании каждого tax_id. С помощью данных файлов сформировано наименьшее по включению узлов дерево,

содержащее все идентификаторы из taxid_index.tsv, и для данного дерева составлены файлы:

- children.tsv – информация о связях узла с дочерними (в смысле принадлежности следующему рангу из исследуемых) для каждого узла tax_id в формате списка разделенных табуляцией идентификатора и списка дочерних идентификаторов, разделенных в свою очередь знаками «;»
  <tax_id> <дочерний_узел1>;<дочерний_узел2>;…
- reference.tsv – информация об узлах поддерева каждого узла tax_id, имеющих эталонные последовательности, в формате списка разделенных табуляцией идентификатора и списка упомянутых идентификаторов, разделенных в свою очередь знаками «;» *<tax_id> <узел1>;<узел2>;…*
- rank.tsv – информация о биологическом ранге каждого узла tax_id в формате списка разделенных табуляцией идентификатора и соответствующего ему идентификатора ранга *<tax_id> <ранг>*,
  где идентификаторы ранга принимают значения: «U» – корень, «D» – домен (надцарство), «P» – тип, «C» – класс или класс по Балтимору, «O» – отряд, «F» – семейство, «G» – род, «S» – вид, «-» – прочие.
- names.tsv – информация об уровне патогенности и научном именовании каждого узла tax_id в формате списка разделенных табуляцией идентификатора и соответствующих ему уровня патогенности и именованию
  <tax_id> <уровень_патогенности> <именование>,
  где уровень патогенности представлен римскими цифрами, принимает значения из множества {I, II, III, IV} и указан на основе Классификации биологических агентов, вызывающих болезни человека, по группам патогенности. [16]

Глава 3. Подготовка данных

3.1. Построение обучающих множеств

Одним из этапов применения методов глубинного анализа данных является подготовка данных, включающая в себя создание множеств объектов, представляющих классы, и формирование представлений объектов в едином пространстве признаков.

Для каждого узла B таксономического дерева обучающее множество для идентификации каждого узла a_i из $children(B) = \{a \mid a$ – дочерний узел $B\}$ (множества дочерних узлов B) строится из всех последовательностей s_j принадлежащих $reference(a_i) = \{s$ – эталонная последовательность для узла $r \mid r$ – узел-потомок a_i, существуют эталонные последовательности для узла $r\}$ (множество эталонных последовательностей в поддереве, образуемом узлом a_i как корнем), принимая в качестве идентификатора каждой последовательности s_j из $reference(a_i)$ соответствующее значение tax_id_i вершины a_i. Обучающее множество наполняется случайно выбранными n_j подпоследовательностями заданной длины $l = 600$, обеспечивая значительное покрытие ($dp_j > 0.3$) каждой последовательности s_j. Значение dp_j вычисляется как отношение суммы длин подпоследовательностей к длине самой последовательности $|s_j|$ ($dp_j = l*n_j/|s_j|$).

3.2. Извлечение признаков

Для применения методов машинного обучения важным является формализация модели вычисления признаков входных объектов с целью применения алгоритма обучения и впоследствии алгоритма классификации в пространстве признаков.

Вектор множества признаков обозначается следующим образом:

$$F = \{f_i\}_{i=1..N} \tag{2}$$

Для каждого признака f_i вводится множество допустимых значений:

$$f_i : X \to d_{f_i}, d_{f_i} \subset R \tag{3}$$

Для обеспечения независимости от длины произвольных входных строк, представляющих нуклеотидные последовательности, были рассмотрены следующие две модели извлечения признаков, формирующие наборы фиксированной длины действительных чисел из отрезка [0; 1]:

- Профиль *K*-мер [20] – вероятностное распределение различных непрерывных фрагментов фиксированной длины (*K*) исходной нуклеотидной последовательности;
- Хешированный профиль (*K*, *t*)-мер [21] – каждый из *K*-меров с использованием бинарной матрицы представляется определенным числом *t*-мер (см. рис. 2), что может позволить повысить устойчивость анализа к присущим биологическим данным точечным мутациям [22].

В работе [20] была отмечена необходимость подбора длины *K* при составлении профиля. Применение больших значений может привести к значительному увеличению времени исполнения, но позволит расширить множество признаков для составления более точной модели.

Для подготовки сформированных выборок к дальнейшему анализу предлагается для каждого классификатора рассмотреть как модель профиля 6-мер ($|F| = 4^6 = 4096$ признаков) и хешированного профиля (16, 4)-мер ($|F| = 16 \times 4^4 = 4096$ признаков).

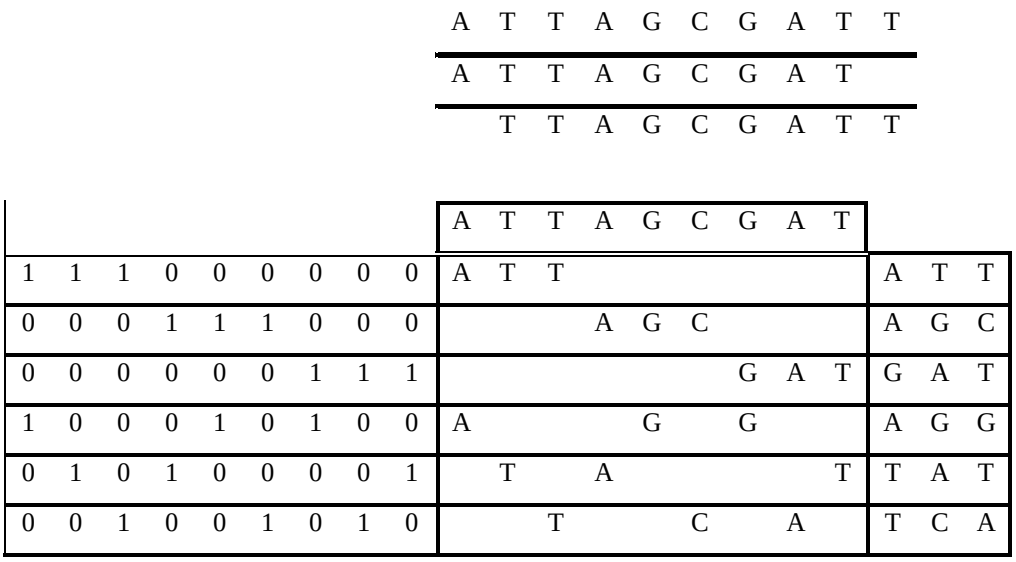

Рис. 2. Пример извлечения (9, 3)-мер

Глава 4. Модель анализа данных

4.1. Случайный лес

В качестве модели классификации предлагается применение метода случайных лесов [23].

Пусть для узла *B* сформировано и переведено в некоторое пространство признаков обучающее множество X_m для классификации среди множества классов $G = \{\text{tax\_id}_i \mid a_i$ принадлежит *children*(*B*)$\}$

Составляются обучающие выборки из подмножеств множества X_m.

Для каждой такой выборки *T* строится дерево решений. На каждой итерации для подмножества *A* (начиная с *A* = *T*) обучающей выборки *T* определяется вектор $\bar{p}_A$ вероятностей p_i выбора элемента с идентификатором *i* (элемента класса *i*) и строится такое разбиение пространства признаков гиперплоскостью $f_k = x_{f_k}$, которое минимизировало бы среднюю меру неоднородности двух полученных подмножеств:

$$\begin{aligned} &\left\langle f_k, x_{f_k} \right\rangle = \arg\min_{f_i \in F, x \in d_{fi}} \frac{1}{2}(\phi(\bar{p}_{A_{fi<x}}) + \phi(\bar{p}_{A_{fi \geq x}})) \\ &f_k \in F, x_{f_k} \in d_{f_k} \end{aligned} \tag{4}$$

Можно заметить, что

$$\frac{1}{2}(\phi(\bar{p}_{A_{fi<x}}) + \phi(\bar{p}_{A_{fi \geq x}})) < \phi(\bar{p}_A) \tag{5}$$

если выбрать в качестве меры неоднородности, например, Gini index:

$$\phi(\bar{p}_A) = \sum_{i=1}^{M} p_i(1 - p_i) \tag{6}$$

В результате построения таких деревьев определяется лес, который затем возможно использовать для принятия решений о принадлежности произвольной последовательности *s* заданного алфавита Σ определенной таксономической группе. Для этого необходимо:

1. Вычислить значения вектора множества признаков для объекта *s*;

2. Применить последовательность правил разбиения пространства признаков на подмножества каждого дерева к последовательности *s* для получения результатов отдельных деревьев;
3. Произвести заключение о выборе идентификатора по принципу большинства (голосования) среди полученных результатов деревьев.

Для обнаружения невозможности принятия результата как достоверного сравниваются относительные частоты для двух классов с наибольшим количеством результатов деревьев равных данным классам. Если данные частоты отличаются более чем на $0.1/|G|$, где $|G|$ - число классов в рассмотрении.

Формирование деревьев на различных выборках из исходного множества позволяет алгоритму исключать влияние групп объектов значительно отличающихся от основного множества в заданном пространстве признаков и повышает разнообразие результатов, важное для классификации по принципу голосования.

При поиске разбиения пространства признаков (4) возможно исключить из рассмотрения часть признаков для удаления признаков потенциально не вносящих новую информацию и способных вносить шум в модель, приводя к снижению обобщающей способности алгоритма. [15]

Глава 5. Подбор параметров модели

5.1. Параметры модели

Для каждой модели случайного леса помимо параметров каждого решающего дерева – набора последовательных разбиений пространства признаков, получаемых в ходе обучения, существуют «гиперпараметры» – параметры, характеризующие определенные ограничения на модель в целом и на процесс обучения каждого дерева решений. [15]

Среди таких параметров случайных лесов выделяют:

- Число деревьев – количество элементов в случайном лесе, которое необходимо обучить. При увеличении данного параметра повышается общее время построения системы и время ее работы, но также увеличивается число разнообразных деревьев, участвующих в принятии окончательного решения, что положительно влияет на качество классификации.
- Максимальная глубина деревьев – ограничение на глубину каждого дерева решений, чем больше, тем более полно строится разбиение пространства гиперплоскостями, что увеличивает время работы, но способно значительно повысить качество.
- Число признаков для поиска разбиения – при сокращении числа признаков в пространстве для разбиения каждым отдельным деревом увеличивается разнообразие правил принятия решений в деревьях и возрастает качество классификации.

Невозможно строго выделить определенный подход направленного поиска оптимальных параметров, так как выбор гиперпараметров происходит до обучения модели, и для расчета критериев качества потребуется перестройка модели с последующей оценкой ее точности и обобщающей способности. Таким образом, поиск оптимальных параметров необходимо производить среди фиксированного набора возможных значений параметров.

В текущей задаче классификации применяются значения гиперпараметров:

число признаков для поиска разбиения: $\sqrt{|F|}$ = 64 для всех методов извлечения признаков;

максимальная глубина деревьев не ограничивается.

Поиск оптимального параметра числа деревьев производится для каждой модели среди значений из множества:

{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000}

Для сравнения получаемых классификаторов и поиска оптимального параметра необходимо ввести критерии качества классификации.

5.2. Критерии качества

Для оценки качества и обобщающей способности алгоритмов классификации применяются такие методы как скользящий контроль по *K* блокам (*K*-fold кросс-валидация) [24] и оценка на контрольной выборке.

Скользящий контроль по *K* блокам производит разбиение обучающего множества на *K* непересекающихся подмножеств и формирует *K* различных обучающих и тестовых (контрольную) выборок таким образом, чтобы обучающая выборка состояла из элементов *K*-1 блоков, а контрольная - из элементов не включенного в обучающую выборку блока. Критерий вводится как средняя точность классификации среди контрольных выборок. Под точностью понимается доля верно классифицированных объектов контрольной выборки. Путем минимизации средней ошибки производился поиск оптимальных параметров.

Оценка на контрольной выборке производит подсчет критерия качества для элементов контрольной выборки, состоящей из элементов, не используемых для обучения классификатора. В качестве критерия используют точность классификации.

Применение независимых от применяемых для обучения множеств тестовых выборок предоставляет несмещённую оценку критериев качества. [24]

5.3. Сравнение моделей извлечения признаков

Для сравнения качества работы классификатора при различных применяемых моделях расчета признаков объектов используется оценка точности на контрольных выборках:

1. Выборка объектов, построенная по принципу создания обучающих множеств;
2. Выборка, построенная по принципу создания обучающих множеств с изменениями структуры строк, симулирующими данные получаемые в результате секвенирования.

Вторая выборка формируется с целью проверки влияния методов расчета признаков объектов на устойчивость классификации к изменениям в определяемых секвенатором строк генетической информации организмов принадлежащих к исследуемым таксономическим группам.

Для организмов, принадлежащих к единой таксономической группе, характерно наличие точечных несоответствий – мутаций. Также известно, что методы секвенирования не являются абсолютно точными и могут вносить ошибки в определение отдельных символов читаемых коротких участков генома организма.

Для симуляции подобных несоответствий в исходных эталонных последовательностях производилась замена на случайный нуклеотид из множества {A, T, G, C} каждого символа с вероятностью 0.001 (вероятностью мутации). После формирования множества по принципу построения обучающих множеств на основе полученного набора последовательностей как на эталонных производилось моделирование ошибок секвенирования с помощью замены на случайный нуклеотид из множества {A, T, G, C} каждого символа каждой короткой строки с вероятностью 0.05 (вероятностью ошибки секвенирования). Каждой полученной последовательности затем присваивался таксономический идентификатор на основании взвешенного качества выравнивания на имеющиеся эталонные последовательности. [20]

Глава 6. Реализация системы

6.1. Выбор инструментов реализации

Для реализации системы классификации нуклеотидных последовательностей на основе машинного обучения, необходимой при решении поставленной задачи, требуется система разработки, поддерживающая работу библиотек машинного обучения.

Для реализации подобных алгоритмов анализа данных используются следующие средства математических вычислений:

- MATLAB
- Python
- R

MATLAB

Matlab – это кроссплатформенный пакет программ применяемый для математических вычислений, моделирования различных процессов и программирования. Он ориентирован прежде всего на численные расчеты, работу с матрицами, визуализацию и создание приложений.

Пакет включает в себя множество модулей, добавлящих к проектам специализированный функционал: оптимизация алгоритмов, проектирование и моделирование динамических систем (Simulink), работа с нейронными сетями и нечеткой логикой и взаимодействие с другими программными продуктами.

Значительная часть функций пакета доступна для прочтения и модификации.

Matlab предоставляет богатые возможности графического представления различного рода информации. [25]

Python

Набор модулей для работы с операционной системой позволяет писать кросс-платформенные приложения. В стандартной библиотеке существуют модули для работы с текстовыми кодировками, архивами, модули

сериализации данных. Библиотека NumPy позволяет производить обработку многомерных массивов значительного размера со скоростью близкой к специализированным пакетам. Для применения известных математических алгоритмов используется SciPy использует NumPy и предоставляет доступ к обширному спектру математических алгоритмов. Python и большая часть библиотек предоставляются в исходных кодах. В сравнении с иными открытыми системами, возможно использование и в коммерческих разработках.

В качестве системы исполнения с открытым исходным кодом Jupyter Notebook позволяет комбинировать в одном документе коды программ с возможностью их исполнения, графику, комментарии. [26]

R

R – свободная программная среда вычислений, для которой предоставляется множество специализированных пакетов в том числе пакет машинного обучения для классификации и регрессии caret. [27]

Для работы с R существуют такие графические интерфейсы как среда разработки с открытым исходным кодом RStudio. [28]

В качестве средства разработки системы был выбран интерпретируемый язык Python ввиду того, для него предоставляется хорошо документированная библиотека Scikit-Learn инструментов глубинного анализа данных. [29]

6.2. Реализация

Были реализованы следующие функции:

- RefineTaxonomy
- ExtractFeaturesKmer
- ExtractFeaturesByMask
- GenerateSubstrings
- GenerateNoisySubstrings
- GenerateSet

- GenerateNoisySet
- RandomForestTrain
- RandomForestTrainBatch
- PredictByThreshold
- Classify
- Report

RefineTaxonomy производит формирование файлов children.tsv, reference.tsv, rank.tsv, names.tsv на основе taxid_index.tsv, nodes.dmp, names.dmp. (см. параграф 2.3)

ExtractFeaturesKmer(fasta_string, kmer) извлекает представление строки fasta_string в виде профиля K-мер – вектора длины 4^K, где K = kmer. (см. параграф 3.2)

ExtractFeaturesByMask(fasta_string, mask) извлекает представление строки fasta_string в виде хешированного профиля (K, t)-мер – вектора длины $v \times 4^t$, где mask – бинарный массив (см. таблицу 1), v – количество строк в mask, K - длина строки в mask, t – число ненулевых элементов в строке mask (см. параграф 3.2).

SubstringArray = GenerateSubstrings(s, dp, l) производит случайный выбор подстрок заданной длины l из строки s с покрытием dp и запись их в массив SubstringArray. (см. параграф 3.1)

SubstringArray = GenerateNoisySubstrings(s, dp, l, m_rat, er_rat) производит случайный выбор подстрок заданной длины l из строки s (каждый символ которой с вероятностью m_rat заменен на произвольный нуклеотид) с покрытием dp с последующей заменой каждого символа в подстроке с вероятностью er_rat на произвольный нуклеотид и запись их в массив SubstringArray. Под нуклеотидом понимается элемент множества {A, T, G, C}. (см. параграф 3.1)

GenerateSet(B, dp, l, kmer, mask) осуществляет формирование обучающего множества для классификации по сестринским идентификаторам узла с идентификатором B, используя функции

GenerateSubstrings(s, dp, l), ExtractFeaturesKmer(fasta_string, kmer), ExtractFeaturesByMask(fasta_string, mask). (см. параграф 3.1)

GenerateNoisySet(B, dp, l, m_rat, er_rat, kmer, mask) осуществляет формирование набора тестовых данных, симулирующих наличие точечных мутаций происходящих с вероятностью m_rat и ошибок секвенирования происходящих с вероятностью er_rat, для классификации по дочерним идентификаторам узла с идентификатором B, используя функции GenerateNoisySubstrings(s, dp, l, m_rat, er_rat), ExtractFeaturesKmer(fasta_string, kmer), ExtractFeaturesByMask(fasta_string, mask). (см. параграф 3.1)

RandomForestTrain(B) применяет функцию GridSearchCV библиотеки Scikit-Learn для модели классификации RandomForestClassifier для поиска оптимальных параметров методом 5-fold кросс-валидации и обучения элемента системы классификации (отдельно для каждой модели извлечения признаков), производящего определение таксономической принадлежности произвольной строки одному из дочерних узлов узла с идентификатором B. (см. главу 4)

RandomForestTrainBatch производит запуск функции RandomForestTrain(B) для всех B – идентификаторов узлов обладающих дочерними узлами. (см. главу 4)

PredictByThreshold(RF, X, threshold) применяет функцию predict_proba(X) объекта RF класса RandomForestClassifier для сравнения двух наибольших относительных частот результатов деревьев для вектора признаков объекта X. Если значения отличаются больше чем на threshold, то в качестве результата классификации принимается идентификатор соответствующий наибольшей из частот, иначе сообщается, что дальнейшая идентификация недопустима. (см. главы 1, 4)

Taxid = Classify(B, rank, X) производит классификацию для вектора признаков объекта X, определяя новый идентификатор C более низкого ранга с помощью PredictByThreshold(RF, X, threshold), где RF – объект класса

RandomForestClassifier составленный для узла с идентификатором B и вызывая Classify(C, rank, X), пока ранг узла с идентификатором B не равен rank, или пока PredictByThreshold не вернет сообщение о недопустимости идентификации, или пока у узла с идентификатором B существуют дочерние узлы. Выводом функции является результат классификации – таксономический идентификатор Taxid.

Report(input, output) выводит для результатов классификации коротких последовательностей (ридов) – файла input, содержащего список присвоенных таксономичесих идентификаторов, и сохраняет в файл output отчёт (см. рис. 3) об образце, включающий следующие поля:

1. В первой колонке – процент (с точностью до 0.01%) ридов, отнесенных к данному таксону (5, 6 колонки).
2. Во второй колонке – число ридов, отнесенных к данному таксону.
3. В третьей колонке – число ридов, отнесенных к данному таксону, но ни к одному из нижележащих в иерархии таксонов.
4. В четвертой колонке – ранг таксона: «U» – корень, «D» – домен (надцарство), «P» – тип, «C» – класс или класс по Балтимору, «O» – отряд, «F» – семейство, «G» – род, «S» – вид, «-» – все остальные категории.
5. В пятой колонке – tax_id данного таксона (целое число).
6. В шестой колонке – научное именование таксона и группа патогенности для патогенных вирусов и бактерий (в скобках римскими цифрами).

	A	B	C	D	E	F	G
1	1,00	5375	595	-	root		
2	0,89	4780	3	D	Viruses		
3	0,62	3355	0	-	ssRNA viruses		
4	0,61	3271	0	C	ssRNA negative-strand viruses		
5	0,61	3271	0	O	Mononegavirales		
6	0,61	3271	0	F	Filoviridae		
7	0,61	3270	0	G	Marburgvirus		
8	0,61	3270	3006	S	(I) Marburg marburgvirus		
9	0,04	227	227	-	(I) Lake Victoria marburgvirus - Ci67		
10	0,01	37	37	-	(I) Lake Victoria marburgvirus - Leiden		
11	0,00	1	0	G	(I) Ebolavirus		
12	0,00	1	1	S	(I) Zaire ebolavirus		
13	0,02	84	1	C	ssRNA positive strand viruses, no DNA stage		
14	0,01	71	0	F	Flaviviridae		
15	0,01	69	0	G	Flavivirus		
16	0,01	69	69	S	(II) Zika virus		
17	0,26	1422	1	C	dsDNA viruses, no RNA stage		
18	0,26	1417	0	O	Herpesvirales		
19	0,26	1417	0	F	Herpesviridae		

Рис. 3. Пример отчёта

Глава 7. Результаты

Были созданы обучающие множества для разделения последовательностей на группу вирусов и группу бактерий.

Также были созданы множества для следующих таксономических уровней вирусов:

1. Класс по Балтимору; [30]
2. Семейство;
3. Род;
4. Вид.

Были созданы множества для следующих таксономических уровней бактерий:

1. Класс;
2. Семейство;
3. Род;
4. Вид.

Построены классификаторы на основе метода случайных лесов для каждого представителя указанных таксономических уровней с проведением поиска оптимальных параметров модели классификации методом скользящего контроля по 5 блокам.

Промежуточные результаты сравнения качества (в том числе и для нескольких методов вычисления признаков объекта) представимы в форме графиков зависимости средней точности классификации от параметров модели.

Рассмотрим результаты в частности для классификации по таксономическим идентификаторам из множества Y = {40050, 40051, 40052, 40053, 40054, 40056, 40057, 40058, 40059, 40060, 40061, 40062, 40065, 40067, 104581, 201490, 204269, 306276, 356862, 464979, 1408137, 1428763, 1679172, 1955493}, $|Y|$ = 24 видов принадлежащих роду *Orbivirus*. В результате поиска в пространстве параметров (см. рис. 4) для модели извлечения признаков на основе профиля 6-мер (непрерывная линия) оптимальным является

количество решающих деревьев равное 600 (точность равна 0,9997), а для (16, 4)-мер (пунктирная линия) оптимальное количество решающих деревьев равняется 700 (точность равна 0,9992)

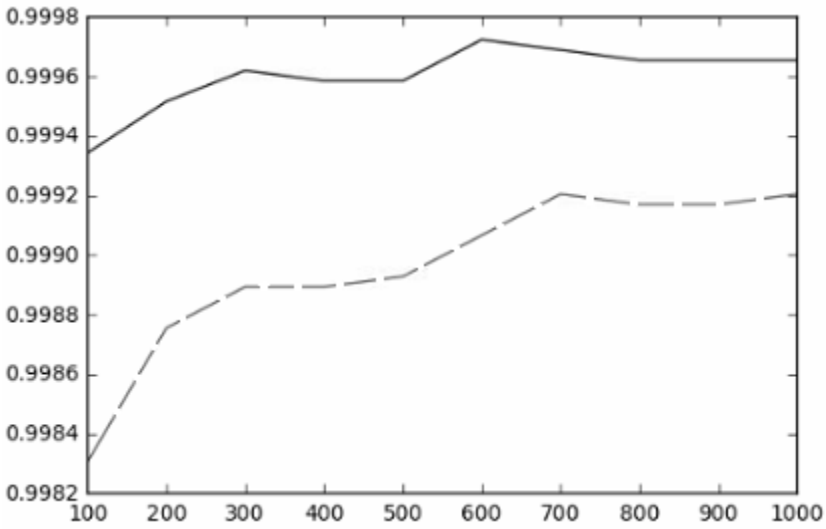

Рис. 4. Поиск оптимальных параметров

В результате сравнения методов как на контрольных выборках подобных обучающим множествам, так и на выборках моделирующих последовательности, получаемые в результате секвенирования образцов, получено большее значение точности для метода 6-мер для каждого из узлов формирования классификаторов. Таким образом, для окончательного составления системы использовались классификаторы на основе метода случайных лесов с моделью расчета признаков на основе профиля 6-мер.

При проверке качества классификации отдельных элементов точность классификации была более 0.98 (0.95 для выборок, моделирующих результаты секвенирования).

Точность работы системы в целом для определения биологических видов составила 0.988.

Время работы на 100000 последовательностях составило 0.3 секунды, что в 10 раз превысило скорость работы BWA-MEM, системы основанной на выравнивании, при запуске на сервере с двумя процессорами Intel® Xeon® Processor E5-2630 с тактовой частотой 2.60 ГГц и оперативной памятью – 64 ГБ

Эффективность работы алгоритма проверена обнаружением всех введенных в образец видов бактерий и вирусов, секвенированных на платформе Illumina NextSeq в лаборатории иммунологии и вирусологии ВИЧ-инфекции Санкт-Петербургского научно-исследовательский института эпидемиологии и микробиологии имени Пастера.

Выводы

Произведено создание базы эталонных последовательностей позволившей реализовать модель классификации в предметной области идентификации патогенных вирусов и бактерий.

Показана методика создания обучающих множеств.

На языке программирования Python реализованы программа подготовки исходных данных, программа обучения элементов системы классификаци, представленных моделью анализа данных RandomForestClassifier, реализованной в рамках библиотеки Scikit-Learn, и программа, производящая определение таксономической принадлежности нуклеотидных последовательностей.

Произведен выбор модели извлечения количественных признаков строковых представлений биологических последовательностей.

Произведена тестовая обработка результатов секвенирования бактерий и вирусов.

Заключение

В процессе проведенной работы получены следующие результаты:

1. Разработан алгоритм биоинформатического анализа идентификации патогенных микроорганизмов.

2. Проведено исследование методов подготовки данных и моделей анализа данных на основе машинного обучения с целью выбора математических, статистических и программных инструментов, необходимых для эффективной реализации разработанного алгоритма.

3. Разработана система определения таксономической принадлежности биологических последовательностей патогенных вирусов и бактерий.

Создание и реализация в результате проведенной работы быстрого алгоритма биоинформатического анализа идентификации патогенных микроорганизмов, включающего в себя методы прикладной математики, статистики и информатики, открывает возможность полномасштабных исследований распространения и мутагенеза опасных для человека микроорганизмов с целью выявления и предотвращения массовых инфекционных заболеваний, что в современных условиях существования общества имеет особенно важное значение.

Выбор и применение интенсивных вычислительных методов на основе выявления общих закономерностей по известным данным, таких, как методы машинного обучения, имеют определяющее значение для достижения цели решения практических задач, возникающих при обработке обширных объемов данных биологических исследований, необходимых для выявления опасных микроорганизмов с целью предотвращения эпидемий. Совершенствование вычислительных методов, баз данных и алгоритмов решения задач анализа биологических данных – это актуальная в настоящее время задача, стоящая перед современной генетикой, эволюционной биологией, вычислительной биологией и другими информационно-ёмкими отраслями фундаментальной биологии.

Выбор иерархической структуры системы классификаторов позволяет проводить точные исследования таксономического состава биологических образцов. Также подобная структура позволяет легко дополнять базу эталонных последовательностей, производя переобучение лишь малого числа элементов системы, находящихся в той ветви таксономического дерева, в которой находятся идентификаторы добавляемых последовательностей.

Разработанный алгоритм идентификации патогенных агентов может применяться как элемент программного конвейера – управляемой последовательности запуска программ-модулей для предоставления информации об оценке таксономической принадлежности с целью ускорения работы и повышения специфичности анализа проводимого другими инструментами биоинформатического анализа.

Разработка системы классификации биологических последовательностей, проведенная в представленной работе с использованием методов машинного обучения, применяющих знания из классических математических дисциплин, методов оптимизации и математической статистики, позволяет быстро и точно производить обработку баз данных биологических исследований с целью обнаружения в кратчайшие сроки возбудителей особо опасных инфекций и, таким образом, значительно ускорять процесс принятия решений о мерах противодействия возможным эпидемиям. Кроме этого, разработанная система классификации может быть предложена для проведения научных исследований современной фундаментальной биологии.

Список цитируемой литературы

1. History of Illumina Sequencing. https://www.illumina.com/science/technology/next-generation-sequencing/illumina-sequencing-history.html.
2. Lavenier, D. PLAST: parallel local alignment search tool for database comparison. BMC Bioinformatics, 2009. 329 p.
3. Li H., Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler Transform // Bioinformatics, 2010. Vol. 26. P. 589-595.
4. Rebecca R., Bede C., Avraam T., David L. R., Mattia P. Challenges in the analysis of viral metagenomes // Virus Evolution, 2016. Vol. 2. Issue 2. P. 427–439.
5. Yu N.; Yu Z. et al. A Comprehensive Review of Emerging Computational Methods for Gene Identification // Journal of Information Processing Systems, 2016. Vol. 12. Issue 1. P. 1-34.
6. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, 2009. 746 p.
7. Smith T. F., Waterman M. S. Identification of Common Molecular Subsequences // Journal of Molecular Biology, 1981. 147. P. 195-197.
8. Federhen S. The NCBI Taxonomy Database // Nucleic Acids Research 2012, Vol. 40. P. 136–143.
9. Pennisi E. Modernizing the tree of life // Science, 2003. 300. P. 1692-1697.
10. Ghosh T. S., Gajjalla P., Mohammed M. H., Mande S. S. A Hidden Markov Model based algorithm for taxonomic classification of 16S rRNA gene sequences // Genomics, 2012. Vol. 99 P. 195–201.
11. Yang B. S., Di X., Han T. Random forests classifier for machine fault diagnosis // Journal of Mechanical Science and Technology, 2008. 22 (9) P. 1716–1725.
12. Sheik M. A., Niranjan J., George B., Vanajakshi L. Application of random forest algorithm to classify vehicles detected by a multiple inductive loop system // 15th International IEEE Conference on Intelligent Transportation Systems, 2012. P. 491-495.

13. Chile P. Progress in Pattern Recognition, Image Analysis, Computer Vision. CIARP, 2011. 572 p.
14. Kumar S., Krishnani K. K., Bhushan B., Brahmane M. P. Metagenomics: Retrospect and Prospects in High Throughput Age // Biotechnology Research International, 2015. P. 121-133.
15. Чубукова И. А. Data Mining: учебное пособие. М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. 382 с.
16. Классификации биологических агентов, вызывающих болезни человека, по группам патогенности. Приложение 3 к СП 1.3.3118-13.
17. O'Leary N., Wright M., Brister J., et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation // Nucleic Acids Research. 2016. Vol. 44. P. 733-745.
18. Lipman D. J., Pearson W. R. Rapid and sensitive protein similarity searches // Science, 1985. Vol. 227 (4693) P. 1435–1445.
19. Kans J. Entrez Direct: E-utilities on the UNIX Command Line. Bethesda (MD): National Center for Biotechnology Information (US), 2013. https://www.ncbi.nlm.nih.gov/books/NBK179288/
20. Aggarwala V., Voight B. F. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome // Nature Genetics, 2016. 48 (4) P. 349–355.
21. Buhler J. Efficient large-scale sequence comparison by locality-sensitive hashing // Bioinformatics, 2001. 17 (5) P. 419–429.
22. Ичас М. Биологический код. — М.: Мир, 1971. 352 с.
23. Brieiman L. Random forests // Mach. Learn., 2001. 45. P. 5–32.
24. Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов. Математические вопросы кибернетики / Под ред. О. Б. Лупанов. М.: Физматлит, 2004. Т. 13, С. 5–36.
25. Иглин С. П. Теория вероятностей и математическая статистика на базе MATLAB. Харьков: НТУ ХПИ, 2006. 612 с.

26. Granger B. E., Perez F. IPython: A System for Interactive Scientific Computing // Computing in Science and Engineering, 2007. Vol. 9, No. 3, P. 21-29.
27. Williams C. K., Engelhardt A., Cooper T., Mayer Z., Ziem A., Scrucca L., Tang Y., Candan C., Kuhn M. M., Package caret, 2015. http://CRAN.R-project.org/package=caret
28. Mark P. J., Jonge E. Learning RStudio for R Statistical Computing. Packt Publishing, 2012. 126 p.
29. Documentation of scikit-learn. http://scikit-learn.org/stable/documentation.html
30. Baltimore D. Expression of animal virus genomes // Bacteriol Rev, 1971. Vol. 35. No 3. P. 235–241.